CLAIMS

Having thus described our invention, what we claim as new and desire to secure by Letters Patent is as follows:

1. A method of outlier detection comprising the steps of:

reducing an outlier detection problem to that of a classification learning problem using unlabeled normal data as positive examples and randomly generated synthesized examples as negative examples, and

then selectively sampling normal and synthesized examples based on uncertainty of prediction to further reduce an amount of data required for data analysis, resulting in enhanced predictive performance while minimizing computational resources including storage requirements.

2. The method of outlier detection recited in claim 1, wherein the step of selectively sampling includes the step of setting sampling probability equal to a measure of uncertainty of prediction for the example.

3. The method of outlier detection recited in claim 2, wherein the measure of uncertainty of prediction is binomial.

4. The method of outlier detection recited in claim 2, wherein the measure of uncertainty of prediction is Gaussian.

5. The method of outlier detection recited in claim 1, wherein the step of selectively sampling includes the step of setting sampling probability proportional to a product of a measure of uncertaint6y and a measure of cust of mis-classifying the same example.

6. The method of outlier detection recited in claim 5, wherein the measure of uncertainty is binomial.

7. The method of outlier detection recited in claim 5, wherein the measure of uncertainty is Gaussian.

8. The method of outlier detection recited in claim 5, wherein the measure of cost is determined as a relative cost of mis-classifying the example in the training data.

9. The method of outlier detection recited in claim 1, wherein the classification learning problem employs an arbitrary algorithm for classification.

10. The method of outlier detection recited in claim 9, wherein the arbitrary algorithm for classification is selected from the group consisting of decision tree learning algorithms, naïve Bayes method, logistic regression method and neural network training algorithms.

11. The method of outlier detection recited in claim 1, further comprising the steps of:

reading a storing normal data at T-real;

generating and storing synthesized data as T-syn; and

wherein the step of selective sampling is performed on data $T := T\text{-real} \cup T\text{-syn}$.

12. The method of outlier detection recited in claim 11, wherein the step of selective sampling uses an underlying, arbitrary classification learning algorithm and proceeds iteratively.

13. The method of outlier detection recited in claim 12, wherein each iteration comprises the steps of:

selecting a smaller sub-sample from the input data;

training of the underlying classification algorithm with the selected data; and

storing a classifier output by the classification algorithm.

14. The method of outlier detection recited in claim 13, wherein the step of selecting is done by choosing examples that are harder to classify with the classifiers obtained in preceding iterations.

15. The method of outlier detection recited in claim 13, further comprising the step of outputting an output hypothesis as a voting function of classifiers obtained in the iterations.

16. The method of outlier detection recited in claim 13, wherein the step of selecting is done by choosing each example with a sampling probability which is set equal to a measure of uncertainty of predicting a label of that example by a collection of hypotheses obtained by calls to the classification algorithm in earlier iterations.

17. The method of outlier detection recited in claim 13, wherein the step of selecting is done by choosing each example with a sampling probability which is set proportional to a product of a measure of uncertainty of predicting a

label of that example by a collection of hypotheses obtained by calls to the classification algorithm in earlier iterations and a measure of cost of misclassifying the same example.

18. A data processing system for outlier detection comprising:

a top control module controlling overall control flow, making use of various sub-components of the system;

a learning algorithm storage module storing a representation of an algorithm for classification learning;

a model output module storing models obtained as a result of applying the learning algorithm stored in learning algorithm storage module to training data and outputting a final model by aggregating these models; and

a selective sampling module accessing data stored in a data storage module, selectively sampling a relatively small subset of the data, and passing the obtained sub-sample to the top control module.

19. The data processing system for outlier detection recited in claim 18, wherein the learning algorithm storage module stores an arbitrary algorithm for classification.

20. The data processing system for outlier detection recited in claim 19, wherein the arbitrary algorithm for classification is selected from the group consisting of decision tree learning algorithms, naïve Bayes method, logistic regression method and neural network training algorithms.

21. The data processing system for outlier detection recited in claim 18, wherein the data storage module comprises two separate modules, one for

storing real data corresponding to "normal" data, and the other for storing synthesized data corresponding to "abnormal" data.

22. The data processing system for outlier detection recited in claim 18, wherein the data storage module comprises a single data storage module providing two logical data storage modules in a single physical data storage module, one for storing real data corresponding to "normal" data, and the other for storing synthesized data corresponding to "abnormal" data.